Paper review: An overview on microarray technologies

Rohmatul Fajriyah

Department of Statistics, Universitas Islam Indonesia, Yogyakarta, 55584 Indonesia

*Corresponding E-mail: rfajriyah@uii.ac.id

ARTICLE INFO

ABSTRACT

How to cite this article:
Fajriyah, R. (2021). Paper review: An overview on microarray technologies. *Bulletin of Applied Mathematics and Mathematics Education*, *1*(1), 21-30.

Bioinformatics is a branch in Statistics which is still unpopular among statistics students in Indonesia. Bioinformatics research used microarray technology, because data is available through to microarray experiment on tissue sample at hand. Microarray technology has been widely used to provide data for bioinformatics research, since it was first introduced in late 1990, particularly in life sciences and biotechnology research. The emergence and development of the Covid-19 disease further reinforces the need to understand bioinformatics and its technology. There are two of the most advance platforms in microarray technology, namely, are the Affymetrix GeneChip and Illumina BeadArray. This paper aims to give an overview about microarray technology on the two platforms and the advantage of using them on bioinformatics research.

Introduction

Bioinformatics is an emerging research field in Indonesia, which currently is un popular yet among statistics students. Bartlett et al. (2017) said that one future of bioinformatics is: "bioinformatics as an academic discipline, which entails departments, under graduate and postgraduate courses, professional chairs and other structural elements of an established discipline". According to the Decree of Director General of Students and Learning Affairs of the Indonesian Ministry of Research Technology and Higher Education Number 46/B/HK/2019, bioinformatics belongs to multi, inter, or trans-discipline field. Students will learn statistics, mathematics, computer/information technology, chemistry, physics, biology, pharmacy, medicine, agriculture, and forestry (Belmawa, 2019).

Bioinformatics covers genomics, transcriptomics, proteomics, metabolomics and translational bioinformatics. It uses genomics technology to produce data, where statistician will be greatly involved. Genomics technologies have developed very fast in recent years from microarrays to the next generation sequencing (NGS). The NGS costs decline significantly and it becomes a favor many researchers. But still there are many researchers using the microarray in their research and the reasons have been explained by Bakers (2013). The application and development of genomic technologies, i.e. microarray, in Indonesia seem rather slow than Singapore and Malaysia let alone in comparison to Europe and US.

On the other hand, it is widely recognized that this technology is very useful and really supports research in medics, molecular biology, and life sciences in general. For instances, gene mapping in cancer, malaria, and other disease research. Recently handling of Covid-19 is an example of how bioinformatics plays a role in finding the necessary drugs. According to Mohs and Greig (2017), Horizny (2019), and Lansdowne (2020) the drug discovery and development process usually could take time 15-25 years, yet we all witnessed that the covid-19 vaccine was developed less than 3 years. This is one of the examples how important and powerful bioinformatics in drug and development research.

In general, the bioinformatics research is conducted by sampling the biological tissue of patient where it is made up of cells. Cells are the fundamental working units of every living organism, where all the instructions needed to direct their activities are contained within the chemical nucleic acid. Nucleic acid is made up of nucleotides. It consists of a nitrogenous base, sugar (pentose) and phosphate. There are two kinds of nucleic acid, deoxyribose nucleic acid (DNA) and ribose nucleic acid (RNA). There are five types of base, namely, adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U). In nucleic acid, the pentose sugars are deoxyribose and ribose.

A DNA is a nucleic acid with sugar component of deoxyribose and the base components are A, C, G and T. The deoxyribose sugar consists of 5 carbons and an oxygen in a ring, and the carbons are numbered 5', 4', 3', 2', and 1'. The ' is read as prime, a naming convention that means to the carbons in the deoxyribose ring, not the carbons of the base.

A DNA takes the form of double helix with two nucleotide chains, with a linear backbone of sugar (S) and phosphate (P). In this form, the direction of the nucleotides in one strand is opposite to their direction in the other strand. The ends of DNA strands are called the 5 'and 3 'ends. It refers to the locations of carbons on the pentose sugars. The structure of DNA can be seen at Figure 1. The double helix is formed due to the hydrogen bonding between base pairs. The bases on one strand are paired with the bases on another strand, according to Watson- Crick base pairing rules, where A specifically pairs with T, and C will pair with G (Amaratunga & Cabrera, 2004; Draghici, 2003; Zhang, 2006). It is repeated millions or billions of times throughout a genome. The particular order of As, Ts, Cs, and Gs dictates whether an organism is human or another species, for example yeast, rice, or fruit fly.
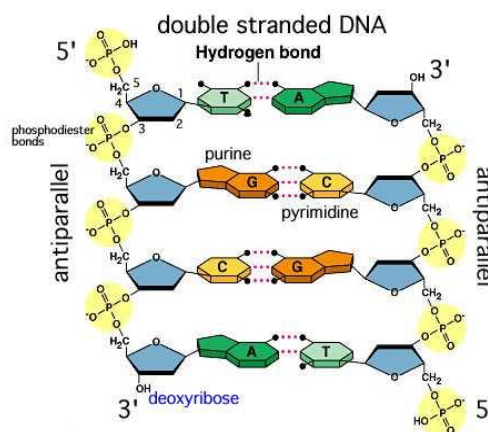

Figure 1. DNA structure (Fajriyah, 2014)

DNA in each human cell is packaged into 46 chromosomes and arranged into 23 pairs. Each

chromosome is a physically separate molecule of DNA that ranges in length from about 50 million to 250 million base pairs (Amaratunga & Cabrera, 2004). Each chromosome contains many genes, the basic physical and functional units of heredity for an organism (Lee, 2006). The structure at the end of chromosome where it is an area of highly repetitive DNA sequences is called telomere, see Figure 2 and Figure 3. An RNA is a nucleic acid with the sugar component of ribose (has an -OH at the 2 'C position, where the DNA sugar has an -H at that position) and the base component is containing base uracil instead of thymine. It is a single stranded. Genes are specific sequences of bases that encode instructions on how to make proteins or an RNA molecule (Amaratunga & Cabrera, 2004; Draghici, 2003; Lee, 2006; Zhang, 2006) The process whereby a gene transfers its genetic code information from DNA into protein, is called a gene expression.
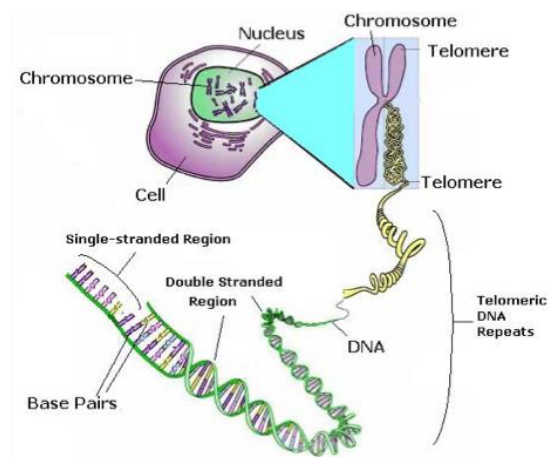

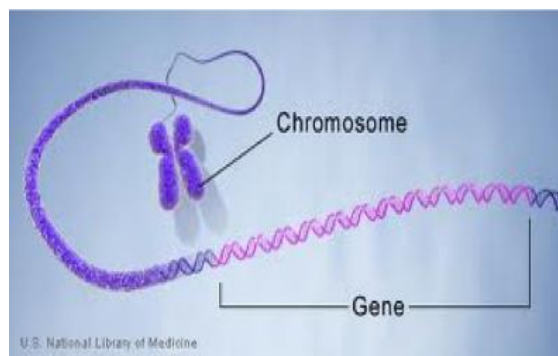**Figure 2.** Cell, Chromosome, and DNA (Burgoine, n.d.)

Figure 3. Gene and DNA (Fajriyah, 2014)

Firstly, the DNA double helix is splitting and developed a condition where one strand of the DNA acts as a template of where the complementary of messenger ribonucleic acid (mRNA) is formed. The mRNA strand then separates. The sequence bases of mRNA are then converted into proteins through the translation step. All the process are formulated in a central dogma of molecular biology (Amaratunga & Cabrera, 2004).

In monitoring the genes expressions, initially, researchers did it sequentially. But in the past few decades, some advances developments have been made through a DNA microarray technology. This technology offers and provides the new paradigm in how the research, techniques and knowledge in life sciences and biotechnology (including genetics engineering, genomics, proteomics and bioinformatics) would be conducted, presented, and used.

DNA Microarray Technology

The DNA microarray is a research tool to investigate the DNA of individual sample or tissue(s) in parallel. The technology provides different methods in fabrication, colours (channels) and types (platforms). For the platforms, in general they are divided into whether they are in a well-arranged surface or on coded beads. The DNA microarray technology could be use for, for instance, genotyping, gene expression, and quantitative protein profiling, in research and clinical purpose.

In molecular biology, A DNA microarray technology used to monitor gene expression in parallel. Gabig and Wegrzyn (2001) define the technology as high-density arrays of DNA or oligonucleotide sequence, known as probes, in thousands of features. These probes will hybridize the mRNA samples in Watson-Crick base pairing. Because there are probes for each gene, this enables us to measure the activity level of genes in a particular sample.

The cells in a human body contain identical genetic material, but the same genes are not active in every cell. To determine which genes are turned on and which are turned off in a given cell, a researcher needs to conduct of microarray experiments, as follows.

1. The mRNA molecules that are present in that cell are extracted, isolated and purified by following the specific protocol, based on the company of the chosen technology.
2. The researcher labels each mRNA molecule using a reverse transcriptase enzyme (RT). This process will generate a complementary oligonucleotide to the mRNA. During that process fluorescent nucleotides are attached to the mRNA.
3. The researcher places the labeled mRNAs onto a DNA microarray slide. The labeled mRNAs that represent mRNAs in the cell will then hybridize -or bind- to their synthetic complementary DNA or oligonucleotides attached on the microarray slide.
4. The researcher then uses a scanner to measure the fluorescent intensity for each feature on the microarray slide.

If a particular gene is very active on a given cell, it produces many molecules of mRNA. Therefore, the hybridization process will generate very bright fluorescence. Genes that are less active produce fewer mRNAs and will produce dimmer fluorescence. If there is no fluorescence, none of the messenger molecules has hybridized to the target on the microarray slide, indicating that the gene is inactive. The gene expression is measured by the intensity value of the scanned image of microarray slide after the steps of hybridization, washing and staining.

Amaratunga and Cabrera (2004), Draghici (2003), Lee (2006), and Zhang (2006) explain that the application of microarray technology is related to the post-genomics era, since a GeneChip contains tens of thousands of probes. Because of that, a microarray experiment can monitor the expression pattern of many genes in parallel and therefore researchers can simultaneously investigate many genes and their interaction at once. Previously, this was not possible: the researcher could only monitor a few genes in one experiment.

The next section will describe the representation of two most popular platforms to measure the gene expression.

The Two Most Used Platforms

Affymetrix

Affymetrix was the pioneer in inventing the technology to produce the high-density oligonucleotide arrays (Amaratunga & Cabrera, 2004). In Affymetrix arrays, also known as high-density oligonu- cleotide arrays or oligonucleotide arrays, a gene is represented by a set of 11-20

pairs of Perfect Match and Mismatch oligonucleotides. Perfect Match (PM) is a short oligonucleotide of 25 bases, corresponding to a specific transcript (Sartor, Medvedovic, and Aronow, 2003). Mismatch (MM) is the same as PM, except that the 13th base element is the complementary base to the central base PM. Affymetrix refers to each PM-MM pair as a probe pair and the entire set of probe pairs for a gene is called a probe set (Amaratunga & Cabrera, 2004). Figure 4 describes the design of Affymetrix probe sets.

Affymetrix used photo-lithographic method to produce the GeneChip. This method is applied on a silica substrate where light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array (Peae et al., 1994). Each applicable probe is selectively "unmasked" prior to bathing the array in a solution of a single nucleotide, then a masking reaction takes place and the next set of probes is unmasked in preparation for a different nucleotide exposure. This process is repeated, until the sequences of every probe become fully constructed and becomes a GeneChip. Usually, it takes no more than 100 repeats.
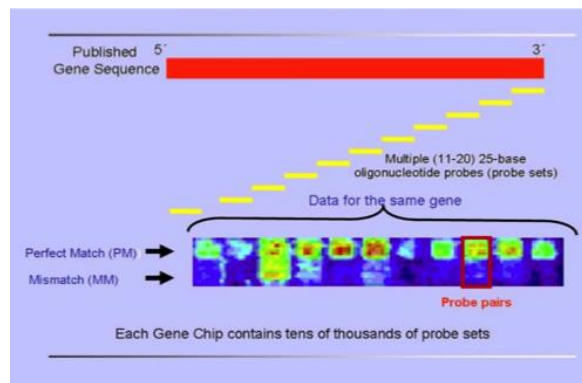

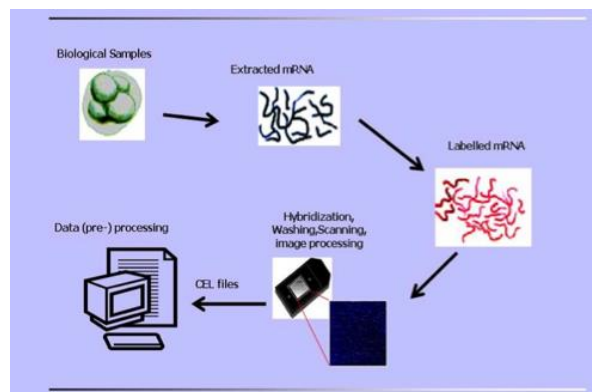**Figure 4.** Affymetrix probe set design (Fajriyah, 2014)

Figure 5. Diagram how Affymetrix GeneChip work (Fajriyah, 2014)

A GeneChip is used to measure the expression measures of the genes from the samples, by doing microarray experiment. The process is similar to that described in Section 2. Figure 5 is a schematic diagram showing the stages of the process in using an Affymetrix GeneChip.

Illumina

Illumina technology is one of the most advance technologies in analyzing gene expression by using microarrays. It has small feature size, dense features and the ability to analyze multiple samples in parallel. Table 1 describes the comparison between Affymetrix and Illumina platforms.

Table 1. Comparison of Affymetrix and Illumina platforms

Aspects	Affymetrix	Illumina
1. Fabrication	By in situ synthesis of 25mer oligonucleotides.	The process involves standard oligonucleotide synthesis methods as for spotted long oligonucleotides arrays. The oligonucleotides are attached to microbeads and then randomly self-assembly in microwells on either of two substrates: fiber optic bundles or planar silica slides.
2. Probe Selection	Uses multiple probes for each gene where one base mismatch probes is developed as controls for nonspecific hybridization.	
3. Design Procedure	Each eligonucleotide is made from one probe and in certain layout of array. The location of each probe has been pre-defined.	Randomly ordered arrays. Each array is assembled on optical imaging, fibre bundle. It is consisting of 50.000 fiber bundle (beads, approximately 3 lijm in diameter and spaced approximately 5 lijm apart) and fused together into hezagonally shaped matrix. Each bead is coated by ten to hundreds of thousands the same probe and each probe is located through a 'decoding' step. Throught the array it provides of 20-40 copies of the same oligonucleotide. Each bead is covered with hundreds of thosand of copies of a specific oligonucleotide that act as the capture sequences in one Illumina arrays.
4. Packaging	Regarding the current packaging, the hybridization and other steps are processed separately.	Are placed on the same physical substrate. Therefore, the hybridization and other steps are performed in a parallel manner.
5. Price	Expensive	The cheapest one among the existing platforms.

Sources: Barnes et al., 2005; Oliphant et al., 2002.

Illumina provides two formats of microarrays, i.e. Fan et al. (2005; 2006) and Steemers and Gunderson (2005): (1) the SentrixR Array Matrix (SAM), and (2) the Sentrix BeadChip (SBC), as illustrated in Figure 6. In Figure 7, the bead is shown to be coated by one oligonucleotide only. In the real bead, it is coated by hundreds of thousands of copies of a specific oligonucleotide.

The Array Matrix arranges fiber optic bundles, each containing 50,000 5-μm fibers, into an Array of Arrays™ format that is compatible with and can access the wells of a 96-well microtiter plate. The fiber optic bundles in the Array Matrix assembly are polished flat on both ends. On one end, the core of each fiber is etched to form a nano well that will accept 3-μm silica beads, each of which has been coated with several hundred thousand oligonucleotides of a particular sequence.
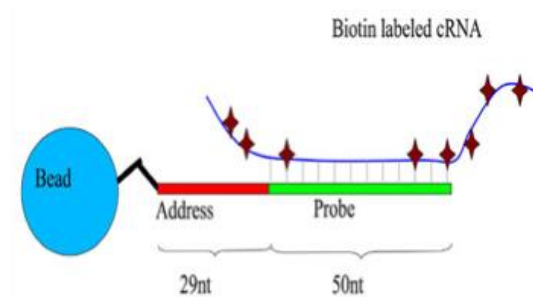
Figure 6. Illumina platforms [8]

Figure 7. The design of an Illumina bead (Grigoryev, 2011)

In the BeadChip format, one to several microarrays are arranged on silicon slides that have been processed by micro-electromechanical systems (MEMS) technology to also have nano wells that support self-assembly of beads.

Steemers and Gunderson (2005) explain the three parts of Illumina BeadArrays manufacturing processes as follow:

1. The first part is the creation of a master bead pool consisting of 1,536-250,000 different bead types. For quality control, it includes the negative control beads. Oligonucleotide capture probes are immobilized individually by bead type in a bulk process. Each bead type in an array comes from a single immobilization event, reducing array-to-array feature variability. The design of the Illumina bead can be seen in Figure 7.
2. The second step is the random self-assembly of the master pool of bead types into etched wells on the array substrate, where each bead type has an average 30 times - a strategy that provides the statistical accuracy of multiple measurements.
3. The third step is the identification of each bead on the array, through a decoding process. This process provides information of each bead and performs a quality control of the feature in every array.

Conclusion

In this paper, we have introduced the two most used microarray technologies, which have been used to generate gene expression data. Further use of these technologies will need to resolve some issues such as stated in these papers (Amaratunga & Cabrera, 2004; Draghici, 2003; Hoheisel,

2006; MacLachlan, Do, & Ambroise, 2004; Lee, 2006; Zhang, 2006). The two of them are: (1) there are the noise and variation contributions from each step of microarray fabrication, (2) once the microarray data are available, the storage, analysis and interpretation of these data present a major challenge due to the massive amount of data generated.

For both issues we could propose statistical modeling and methods to provide the accurate, precise and reliable conclusion of microarray data at hand. The implementation of pre-processing methods and models would solve the first issue (see Fajriyah, 2014; 2015; 2016). On the other hand, the big data techniques could rightly handle such kind of data on the second issue.

The author wishes that in the near future, Indonesia will catch-up in using this technology to analyze data many more than today. The benefit of this technology is, we could investigate thousand of genes in parallel in one experiment. Furthermore, we could invent data analysis method which is more appropriate for the type of data at hand, in which the statisticians definitely will contribute significantly.

References

Amaratunga, A., & Cabrera, J. (2004). *Exploration and analysis of DNA microarray and protein array data*. New Jersey: John Wiley & Sons.

Bakers, S. C. (2013). Next-generation sequencing vs. microarrays: Is it time to switch?. *Genetic Engineering & Biotechnology News*. Retrieved March 4th, 2021 from http://www.genengnews.com/gen-articles/next-generation-sequencing-vs-microarrays/4689/.

Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., & Pavlidis, P. (2005). Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Research*, *33*(18), 5914-5923.

Bartlett, A., Penders, B., & Lewis, J. (2017). Bioinformatics: indispensable, yet hidden in plain sight?. *BMC Bioinformatics*, *18*(1), 1-4.

Belmawa. (2019). *Decree of Director General of Students and Learning Affairs of the Indonesian Ministry of Research Technology and Higher Education Number 46/B/HK/2019 about the list of department in higher education*. Jakarta: Ministry of Research Technology and Higher Education.

Burgoine, K. (n.d.). *Origins of G-Quadruplex DNA*. Retrieved February 22th, 2021 from https://www.ch.ic.ac.uk/local/projects/burgoine/origins.txt.html.

Draghici, S. (2003). *Data analysis tools for DNA microarrays*. Florida: Chapman & Hall.

Fajriyah, R. (2014). Microarray data analysis: Background correction and differentially expressed genes. *Dissertation*. Styria: TU Graz.

Fajriyah, R. (2015). Generalized beta convolution model of the true intensity for the Illumina BeadArrays, *Thailand Statistician*, *13*(2), 145-167.

Fajriyah, R. (2016). A study of convolution models for background correction of BeadArrays, *Austrian Journal of Statistics*, *45*(2), 15-33.

Fan, J. B., Hu, S. X., Craumer, W. C., & Barker, D. L. (2005). BeadArrayâĎć-based solutions for enabling the promise of pharmacogenomics. *BioTechniques*, *39*, 583-588.

Fan, J. B., Gunderson, K. L., Bibikova, M., Yeakley, J. M., Chen, J., Garcia, E. W., Lebruska, L. L., Laurent, M., Shen, R., & Barker, D. (2006). Illumina universal bead arrays. *Methods in Enzymology*, *410*, 57-73.

Gabig, M., & Wegrzyn, G. (2001). An introduction to DNA Chips: principles, technology, applications and analysis. *Acta Biochimica Polonica*, *48*(3), 615-622.

Grigoryev, Y. (2011). *How DNA microarrays are built*. Retrieved February 22nd, 2021 from https://bitesizebio.com/7463/how-dna-microarrays-are-built/.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nature, 7*, 200-210.

Horizny, C. (2019). *The drug discovery process*. Retrieved February 22nd, 2021 from https://www.taconic.com/taconic-insights/quality/drug-development-process.html.

Lansdowne, L. E. (2020). *Exploring the drug development process*. Retrieved February 22nd, 2021 from https://www.technologynetworks.com/drug-discovery/articles/exploring-the-drug-development-process-331894.

Lee, M. L. (2006). *Analysis of microarray gene expression data*. New York: Springer.

McLachlan, G. J., Do, K. A., & Ambroise, C. (2004). *Analyzing microarray gene expression data*. New Jersey: John Wiley & Sons.

Mohs, R. C., & Greig, N. H. (2017). Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 3*(4), 651-657. doi: 10.1016/j.trci.2017.10.005.

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., & Chee, M. S. (2002). BeadArray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTechniques, 32*, S56-S61.

Peae A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., & Fodor, S. A. (1994). Light-generated oligonu- cleotide arrays for rapid DNA sequence analysis. *Proceeding National Academy of Sciences, 91*, 5022-5026.

Sartor, M. A., Medvedovic, M., & Aronow, B. J. (2003). Microarray data normalization: The art and science of overcoming technical variance to maximize the detection of biologic variance. In E. Blalock (Ed.), *A Beginner's Guide to Microarrays*. Massachusetts: Kluwer Academy Publishers.

Steemers, F. J., & Gunderson, K. L. (2005). *Pharmacogenomics, 6*, 777 - 782.

Zhang, A. (2006). *Advanced analysis of gene expression microarray data*. Singapore: World Scientific Publishing.

This page is intentionnaly left blank.