Genomes, Transcriptomes, Proteomes, Metabolomes in Bioinformatics Investigating Alzheimer’s Disease

Conghan Li

fisheries college, Ocean university of China, Qingdao, China

liconghan@stu.ouc.edu.cn

Abstract. Alzheimer’s disease is affecting many people, especially people older than 65, living in the world. However, the cause and cure of these disease continue to confuse the scientists. Omics analysis, which includes four levels: genomes, transcriptomes, proteomes and metabolomes analysis, can dig out the information lying behind large amount of data and it’s really useful for complex disease like AD (Alzheimer’s Disease). In this article, I talked about the four levels of omics analysis investigating AD in details. In each level, the general workflow, the information we can get and the examples in AD have been involved. This paper will give the readers a general view of the function of omics analysis in AD.

Key words: Genomes, transcriptomes, proteomes, metabolomes, Alzheimer’s Disease

1. Introduction

In the last 30 years, the development of molecular biology results in a dramatic expansion of biological information which can be divided into different types, such as molecule sequence data, protein secondary structure and tertiary structure data. Bioinformatics is a subject aiming at processing these data and digging deep into the data for hidden biological secrets. An important part of bioinformatics data is omics data, and in order to explain for these data we conduct omics analysis. Omics data includes genomes, transcriptomes, proteomes and metabolomes data. Each level of data has its own analysis approaches, but there are also many similarities. Moreover, they share the same main idea: make comparisons and the difference shows the answer. AD is a progressive neurologic disorder and it's the most common cause of dementia[1]. It will lead to memory loss, impaired visual spatial ability, personality changes and so on. About 5.8 million people at age 65 or older in America are living with Alzheimer's disease[2]. Alzheimer’s disease is very complex, age, heredity and family history all have to do with its development. Such a kind of complex disease is caused by both genetic and environmental factors. Therefore, it may be quite hard to finger out it from a limited perspective, different level of omics analysis which deal with wide range of and large amount of data may be more useful for such a kind of complex disease[3]. Up to now, although many researches on Alzheimer's disease have been conducted, the cause is still unknown and there is no cure for this disease. Currently, we usually combine drug therapy, non-drug therapy and careful nurse to reduce symptoms and delay disease progression[4]. However, genomics analysis can identify these diseases related variations therefore we can further detect new therapeutic targets; transcriptomes analysis can tell the differential expressed genes which helps us to learn about the pathology; proteomes and metabolomes help us to identify new disease related bio-markers. Considering these advantages, conducting omics analysis can be useful for studying deeply into Alzheimer’s disease. In this paper, I meticulously introduced the differential level of omics analysis and how they can be used in Alzheimer’s disease. In each level, the general workflow, the information we can dig out from the data and the examples of its application in Alzheimer’s disease are talked about. Moreover, as DNA sequencing, an important step in genomes analysis, has a long and attractive history, it’s also introduced.

2. Genomes analysis in ad (alzheimer’s disease)

2.1. Brief developing history of DNA sequencing

DNA sequencing is the basis of genome analysis whose primary sequence data comes from DNA sequencing. DNA sequencing has undergone a long developing history starting from the first-generation techniques including chain termination method and chemical degradation method. (Fig.1) Chain termination method uses ddNTP to terminate the synthesis reaction. After electrophoresis and autoradiography, we can read the sequence. Chemical degradation method utilizes Isotopic to label the end of one side of a DNA chain. Then several different chemicals are used to cut off the DNA chain at specific sites. After electrophoresis and autoradiography, we can get the sequence. The first-generation methods are time-consuming and laborious which are the main reasons why they are rarely used now.

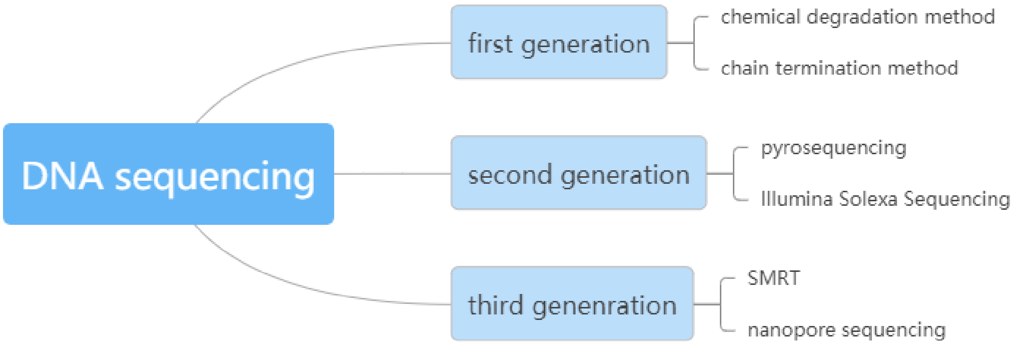

Figure 1. Classification and history of DNA sequencing.

Then came the second generation of sequencing which includes pyrosequencing and illumina sequencing. This generation uses “sequencing by synthesis strategy” and starts the generation of high-throughput sequencing. Pyrosequencing uses sulfurylase. When there is a base added into the chain, the base will release a pyrophosphate which will be converted into ATP by sulfurylase. Then ATP fuels luciferase and finally cause visible light. Illumina sequencing uses fluorescently-labeled nucleotides and they are reversibly terminated so that we can identify each base at each synthesis step. Moreover, bridge amplification strategy amplifies the signal. Compared with illumina sequencing, pyrosequencing has a lower throughput and costs more, so it’s almost abandoned by people now. However, illumina sequencing is still widely used.

In recent years, third generation sequencing method including SMRT (Single Molecule Real Time) and nanopore sequencing have been invented. In SMRT, single molecule DNA polymerase is fixed inside the zero-mode waveguide. Nucleotides were fluorescently labeled. These nucleotides co ming inside and out the waveguide produce a stable background fluorescence. Once a nucleotide is adding into the chain, the significant difference in fluorescence will tell us the information of this added base. As for nanopore sequencing, its sequencing principle does not depend on DNA polymerase which makes it different from other high-throughput sequencing methods. It uses a small nanopore and when a specific base comes across it, the chemical difference in four kinds of base will lead to differential current so that we can distinguish the base. The main strength of third generation is the significantly improved sequencing length which rises the accuracy of assembly, but the shortcoming is that third generation has high error rate.

The comparison of three sequencing generations is shown in Table 1.

Table 1. The comparison of three sequencing generations.

Generation/items	Method	Whether based on PCR reaction	Read length	Strengths	Weaknesses
First generation	Chemical degradation method	No	Up to 1000bp	1. High quality 2. Reads are longer compared to the second generation. 3. Fit for small-scale sequencing for single gene, short genome fragment and re-sequencing.	1. Depend on enzyme so the reads can't be too long. 2. Low-throughput, time-consuming and costly. 3. Need primers, and some complex fragments that are difficult to perform PCR reactions cannot be sequenced.
	Chain termination method	Yes			
Second generation	Pyrosequencing		400 bp	1. Long reads. 2. short running time.	1. Low throughput. 2. Higher cost.
	Illumina sequencing		100-150bp	1. High throughput. 2. Lower cost. 3. Flexible.	1. Short reads make it hard to analysis. 2. Long running time.
Third generation	SMRT		Dozens to hundreds Kb	1. Quickly 2. Long reads. 3. Nanopore directly detect DNA/RNA sequence.	1.High error rate. (Requires self-correction or second-generation data for correction)
	Nanopore sequencing	No	Hundreds Kb to several Mb		

2.2. The general workflow and related techniques used in genomes analysis

Genomes analysis is the first developed one among four levels of omics analysis. As for sequencing, a common strategy is to combine relatively low cost and low error rate illumina sequencing and third generation sequencing which provides convenience for assembly as it gets long reads. The development of high-throughput sequencing (second and third generation sequencing) makes it possible to do genomes analysis because large amount of genomes sequence data is available. However, low-throughput sequencing operation is tedious, time-consuming and laborious, and cannot be automated, which causes it hard to get enough genomes sequence data.

The first step for genomics analysis is genomic sequencing and we can adapt paired-end method and mate-pair method to facilitate the next step: assembly. Reference assembly are adapted when we have reference sequence. If not, we do de novo assembly and the main approaches are Overlap-Layout-Consensus and algorithm based on grapy theory. With these genomic sequencing data, we can make further gene prediction and its functional annotation.

2.3. The information we get from genomic data after analysis

In general, after analyzing the assembled genome sequence data, we can get a lot of information. Firstly, we can annotate unknown genes for a certain species using homology method or de initio method (mainstream approaches are based on hidden Maekov model). Homology method utilizes related species' known genes to do sequence alignment or using de initio method. (Quite similar to the relationship between reference assembly and de novo assembly) Secondly, we can do gene functional annotation to identify gene's function. We can also construct molecular phylogenetic tree to learn about the evolution condition. As for AD, detecting variations including SNP (single nucleotide diversity), Indel and SV (structural variation) and so on is really helpful. We can identify these AD related genomic variations through sequence alignment between AD patients and nondemented control subjects. Based on our findings, we can further detect new therapeutic targets. There's one thing that needs to be pointed out that even for the same piece of genomic data, the amount of information we get varies depending on our analysis ability. Statistical methods and machine learning methods should be combined together for information mining, and we shouldn't stop developing better algorithm[3]. This is true for each level of omics analysis.

2.4. Application of genomes analysis in AD

There are many examples of researchers conducting genomics analysis to learn more about the AD related genes and variations. A group of researchers have found that variant happened near the gene B1N1 was a LOAD (late-onset Alzheimer's disease) risk modifier through GWAS(genome-wide association study) which identify SNP by studying genome of a large group of people and then identify the disease related SNP by alignment between people with and without a certain disease[5, 6]. Another group of researchers developed a machine learning method to identify all the AD related genes across the whole genome and the classifying accuracy is 84.56%. Their work are based on data got from public Alzheimer's disease databases and other databases and helped to expend the spectrum of AD related genes[7]. In another study, the researchers proposed a machine learning model to predict AD according to gene expression and DNA methylation data. They focused on developing the data processing model in genomes analysis. They applied both gene expression and DNA methylation profiles[8]. The prediction model may be applied to the clinic in the future.

3. Transcriptomes analysis in AD

3.1. The general workflow and related techniques used in transcriptomes analysis

The general workflow for transcriptomes analysis is showed in fig.2. As both DNA and RNA are nucleotide chains, the first several steps of transcriptomes analysis are quite similar to genomes analysis. Firstly, RNA-Seq is conducted to get RNA sequence data, then de novo assembly or genome-guided assembly follows. After assembly, a different step from genomes analysis is quantification. As different genes have different expression levels, the amount of expressed RNA can reflect the expression condition of the original genes, which makes quantification of transcripts very meaningful. Before quantification, normalization helps us to standardize the amount of RNA so that we can compare or integrate the sequencing data got from different simple despite the different sequencing depth. Then, differential expression analysis can find the genes expressed differently compared with other samples got from different tissues or different developmental stage. Next, clustering analysis helps us to classify these found RNA. Then we can conduct GSEA (gene set enrichment analysis) which focus on gene sets that include genes with the same function or chromosomal location or regulation[9]. GSEA usually bases on two database: GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes). In the end, we can use co-expression network analysis to learn about the interactions between genes and establish the network between genes with similar functions.

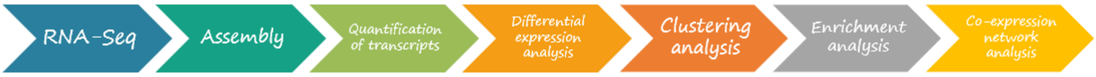
RNA-Seq
Assembly
Quantification of transcripts
Differential expression analysis
Clustering analysis
Enrichment analysis
Co-expression network analysis

Figure 2. workflow of transcriptome analysis.

As for non-coding RNAs, they play a significant role in gene expression regulation and many serious diseases are associated to them, such as AD. Thus, scientists are paying more and more attention to them[10]. As for miRNAs, we can identify them from the transcriptome by calculation according to their characteristics and we can predict a certain miRNA's target gene. For IncRNAs, similar to miRNAs, the main work focusses on their identification and function prediction.

3.2. The information we get from transcriptomes data after analysis

Compared with genomes analysis, the kinds of information we get from transcriptomes data are more diverse. The information lying behind transcriptomes data are the expression condition at the specific time and the specific tissue and the expression regulation condition that related to non-coding RNAs. We can identify the disease related non-coding RNAs and genes which expressed differently in patients compared with normal people. What's more, we can establish disease related gene regulatory networks and pathways.

3.3. Application of transcriptomes analysis in AD

In one study, the researchers constructed gene regulatory network for AD patients and found that specific portions are different from normal people. They also found that the most significant one is immune- and microglia-specific module[11]. That's a typical example of conducting transcriptomes analysis in AD. In another study, scientists did miRNA differential expression analysis between AD patients and nondemented control subjects to find AD associated miRNAs. They found that miRNA-146a amount in cerebrospinal fluid of AD patients is significant low. Maybe they can do further research on miRNA-146a's target gene and its function to explain the reason why it's associated with AD[12].

4. Proteomes analysis in AD

4.1. The general workflow and related techniques used in proteomes analysis

As protein is a completely different substance compared with DNA or RNA, so the first few steps of separation and identification in proteomes analysis is quite different. However, after obtaining the proteomes basic information data, the following analysis methods will be quite similar. In fact, the analysis methods in the whole omics analysis are unified. The high-throughput protein separation technologies involve two-dimensional gel electrophoresis and high-performance liquid chromatography. After separation, we need to identify each protein and the mostly used high-throughput technology mass spectrometry involves MALDI (matrix assisted laser desorption/ionization) and ESI (electro-spray ionization). Just like transcriptomes analysis, we can conduct quantification and differential expression analysis on the separated and identified protein. What's more, we can also do research on PTM (posttranslational modification) and protein interaction.

4.2. The information we get from proteomes data after analysis

Proteomes data shows the gene expression condition at translation level. Trough differential expression analysis, we can identify these disease related genes such as AD. Moreover, the protein, as a disease related biomarker, can be used in disease diagnose. We can also identify protein function, interaction, modification and intracellular localization.

4.3. Application of proteomes analysis in AD

In a study, the researchers used the already identified biomarkers (the identification used proteomes analysis like what I just mentioned above) respectively for AD and subcortical axonal degeneration and use the relationship between these biomarkers to determine that whether these two diseases relate to each other. In this study, results got from proteomes analysis worked as the necessary basis[13]. In another study, scientists found modules of co-expression proteins related to AD based on differential expression analysis. They also fund that RNA bonding proteins and alternatively spliced proteins are abundant in these modules[14].

5. Metabolomes analysis in AD

5.1. The general workflow and related techniques used in metabolomes analysis

The first step is sample preparation and the very first thing to do is to quench all the biochemical processes happening in the sample as many of the metabolite can rapidly degrade or react to produce other substances. The extraction techniques include SPE (Solid-phase extraction) and chromatography[15]. Then the metabolite identification is done by MS (mass spectrometry) or NMR (nuclear magnetic resonance). Then we have to do data processing over the feature matrix got from MS or NMR.

5.2. The information we get from metabolomes data after analysis

From the results of metabolomics information analysis, we can detect new disease-related biomarkers and the biomarkers enable us to quickly diagnose the patients, so that personalized medicine may be possible in the future with the help of metabolomes analysis and researches[16].

5.3. Application of metabolomes analysis in AD

Raúl González-Domínguez and his colleges conducted metabolomes analysis using blood serum samples and identified the differential level of phosphatidylcholines. Thus, it can be regarded as a biomarker for AD[17].

6. Conclusion

Genomes analysis focus on genomes data and as for applying to AD, we can detect related variations, based on this, we can further detect new therapeutic targets. Transcriptomes analysis get information from transcriptomes data. It can detect disease related genes, non-coding RNAs, gene regulatory networks and pathways. Proteomes analysis analyzes the data of all the expressed protein by a cell or tissue. By comparing the data of AD patients and nondemented control subjects, we can detect new AD related biomarkers and related genes. Metabolomes analysis analyzes all the metabolite in a cell or tissue and we can find biomarkers useful for disease detection by this.

The core and unity of each level of omics analysis is to detect important information by comparation between different data. For each level of omics analysis, for the same piece of genomic data, the amount of information we get varies depending on our analysis ability. So we should combine statistical methods and machine learning methods for information mining,[3]and we shouldn’t stop developing better algorithm.

Compared to the study of cancer, the omics study investigating AD is limited, more analysis on bioinformatics data of AD should be conducted. With the development of sequencing technology, a large amount of high-throughput data came out, so we need to develop the analysis approach and algorithm to better deal with these data. Although many AD related genes have been detected, the major genetic contribution to AD remains unknown. Moreover, the direct and specific effects and functions of these genes are not very clear. What’s more, the current findings are not so good adapted to the clinic, the detection and treatment strategy have to be improved.

References

[1] Simon, R.P., M.J. Aminoff, and D.A. Greenberg, *Clinical neurology*. 2009: Lange Medical Books/McGraw-Hill.

[2] Matthews, K.A., et al., Alzheimer's & Dementia. 15(1) 17-24 (2019).

[3] Tan, M.S., et al., Computers in biology and medicine. 139 104947 (2021).

[4] Vaz, M. and S. Silvestre, European Journal of Pharmacology. 887 173554 (2020).

[5] Carrasquillo, M.M., et al., Journal of Alzheimer's Disease. 24(4) 751-758 (2011).

[6] Uffelmann, E., et al., Nature Reviews Methods Primers. 1(1) 1-21 (2021).

[7] Huang, X., et al., BMC neurology. 18(1) 1-8 (2018).

[8] Park, C., J. Ha, and S. Park, Expert Systems with Applications. 140 112873 (2020).

[9] Subramanian, A., et al., Proceedings of the National Academy of Sciences. 102(43) 15545-15550 (2005).

[10] Mattick, J.S. and I.V. Makunin, Human molecular genetics. 15(suppl_1) R17-R29 (2006).

[11] Zhang, B., et al., Cell. 153(3) 707-720 (2013).

[12] Müller, M., et al., Neurobiology of aging. 35(1) 152-158 (2014).

[13] Skillbäck, T., et al., Alzheimer's research & therapy. 5(5) 1-10 (2013).

[14] Johnson, E.C., et al., Molecular neurodegeneration. 13(1) 1-22 (2018).

[15] Villas-Boas, S.G., et al., *Metabolome analysis: an introduction*. 2007: John Wiley & Sons.

[16] Fernie, A.R., Phytochemistry. 68(22-24) 2861-2880 (2007).

[17] González-Domínguez, R., T. García-Barrera, and J.-L. Gómez-Ariza, Chemical Papers. 66(9) 829-835 (2012).