＜シンポジウム 14―1＞神経ゲノミクスの最先端

Whole Exome Sequencing

Karen Nuytemans, Ph.D. and Jeffery M. Vance, M.D., Ph.D.

（臨床神経 2010;50:952-955）

Key words：Next generation sequencing, Whole exome sequencing, Rare variants, Neurogenetics

Over the years there have been many different approaches and techniques that have been utilized to gather genetic information on family and patient data. Early on these focused on using family information and the pattern of inheritance of the disease in the family. The general location of the disease gene on the chromosomes could be determined using linkage analysis. But this required large families and good family histories. Association studies moved away from family data, but still required large numbers of patients sharing the same disease. Over the past several decades we have had continuing growth in the number of genetic "markers" that we could use in linkage and association analyses, the most recent work utilizing the specific sequence changes called single nucleotide polymorphisms or SNPs. But the use of these markers did not eliminate the need for large families or large population sizes. Thus, despite advances in genotyping technology, the parameters of the applications in which these techniques could be used have not changed much in 30 years or so. This prevented many of the most interesting cases and families with Mendelian inheritance (due to a mutation in a single gene) from being studied, as they were too small for analysis, or too rare.

Sequencing genes provided more information, but the previous existing technology was too expensive and labor intensive to do a large number of genes. Thus investigators were dependent on choosing candidate genes for study, which historically has proven to be fairly inefficient.

But this paradigm of needing large families to investigate a disease, or choosing a few candidate genes to test a hypothesis is now changing. The rapidly emerging sequencing technology has now allowed the first practical medical application of the most useful and enlightening measure of genetic information, the DNA sequence itself.

The output of sequencing has been increasing at a rapid pace over the past several years, passing even the growth of the computer chip made famous by "Moore's law" (Moore, G. E. 1965). During the 1990's and early in this century, capillary sequencers, which produced the data for the human genome project, could identify 500 to 1,000 base pairs of sequence at a time. While impressive relative to previous slab gel techniques, the Next Generation Sequencers (NGS) of today have increased that sequencing output a million fold or more (Fig. 1).

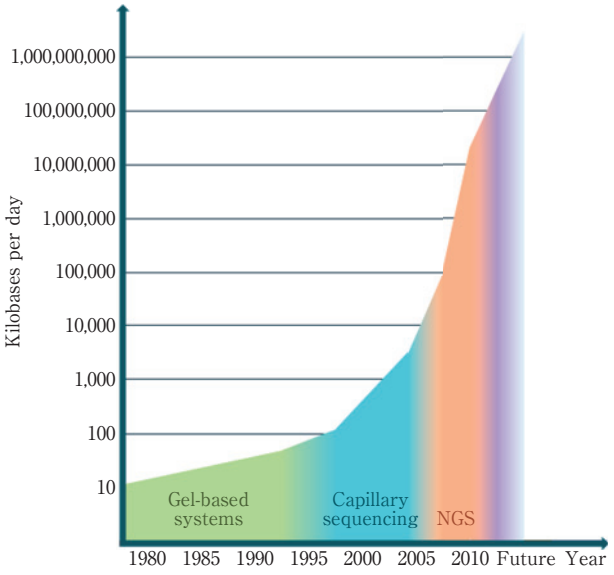

Fig. 1 Improvement in output of sequencing through time.

This also means that the cost of sequencing per base pair has dropped precipitously. The human genome is approximately 3.4 billion base pairs. The cost of sequencing was approximately 10 million dollars in 2003. This dropped to around a million dollars in 2007, less than a $100,000 by 2009, and is expected to be close to the $1,000 goal in 2011-2012.

How has this remarkable increase been accomplished? Sanger sequencing has been the mainstay of sequencing for many years. The Sanger sequencing method that was used for the human genome project utilizes a small primer (usually 20 base pairs), that allows the sequencing reaction to

Department of Human Genetics, University of Miami〔1501 NW 10th Ave, Biomedical Research Building, Suite 616 Miami, FL 33136, USA〕

（Received: 22 May 2010）

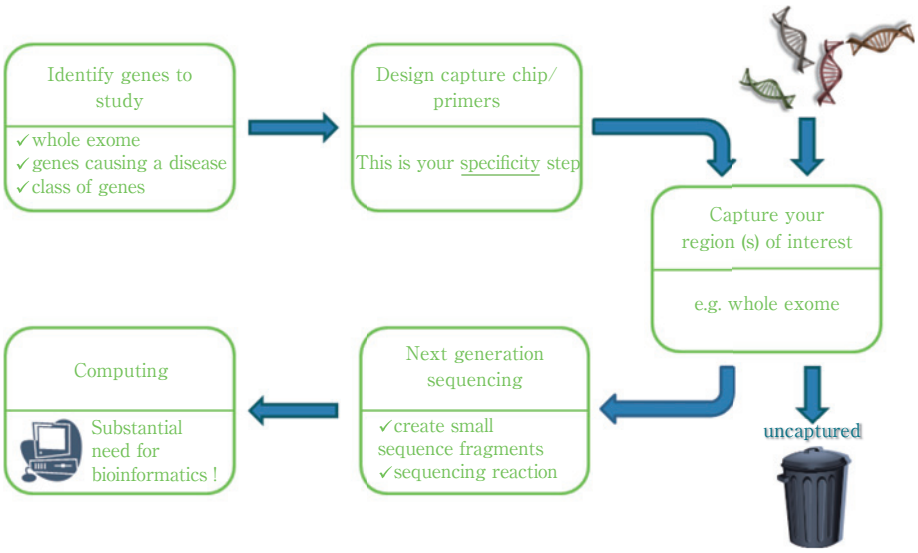

Fig. 2 Next generation sequencing workflow.

start exactly at the location the investigator wishes to begin sequencing. But the use of this primer prevents any significant increase in the amount of sequencing that can be done, as only one primer can be utilized in each sequencing reaction, and any sequencing reaction is only accurate for a limited number of sequences.

The solution to this problem was to eliminate the primer, and create millions of small (50 to 100 base pair) parallel sequencing reactions in a single machine. This approach, called Next Generation Sequencing or NGS, is very powerful, but creates new challenges. First, the complexity and cost of NGS sequencers removes the sequencing labor from the individual laboratory and moves it to a core facility. Second, the massive amount of sequence created now requires a university or corporation to support a large computing and bioinformatic network to handle the data, as the current NGS sequencing reactions produce in the neighborhood of 300 billion base pairs of sequence per run. It is this challenge of handling this amount of data that is likely to be the rate limiting factor of maintaining the current rate of sequencing growth. Third, the sequencing reaction still needs some specificity as to where to start or what regions to sequence, at least until whole genome sequencing is economically practical. As the specificity is no longer in the sequencing reaction itself (the primer is eliminated), a new step to NGS is required, the "capture". Here a template is made of the DNA to be sequenced, and the corresponding DNA of the subject is hybridized to the designed template, to "capture" the DNA for sequencing. This captured sequence is then chopped up, and used for the sequencing reaction (Fig. 2).

So what are the advantages and applications of this new technology? Well, obviously, an enormous amount of sequence can be produced. This allows the sequencing to be much more accurate than Sanger sequencing, and the desired amount of accuracy can be chosen by the user. NGS can sequence very large regions, or a very large number of genes, or even all the exons of all the known genes, which is called whole exome sequencing. At this time, the first generation exome assays or kits capture approximately 92% of the consensus sequence. As our understanding of the human exome changes over time, so will the design of these capture kits.

Whole exome sequencing removes the guess work out of choosing which candidate genes to sequence for a single gene disorder; you sequence almost all of them. Further, as you are sequencing all of the exome, you don't have to have families large enough for linkage analysis, which was previously used to give you a clue where on the genome to sequence. Thus, small single gene (Mendelian) families, previously unable to be analyzed, can now have their genetic defect determined. NGS is now the technique of choice in working up any unknown Mendelian family, and is likely to change the way clinical medicine is practiced. For example, if the cost of sequencing becomes less than many of the traditional analytical tools, such as electromyography and nerve conduction studies in the workup of potentially inherited peripheral neuropathies, it could replace them as part of the initial workup in many situations. Further, large number of genes can now be grouped for sequencing (say, all the genes for Charcot-Marie-Tooth disease), and sequenced for about the cost of one gene previously.

Will this technology also be helpful when studying patients who are diagnosed with a common, complex disease and who do not always show obvious Mendelian inheritance? Well,

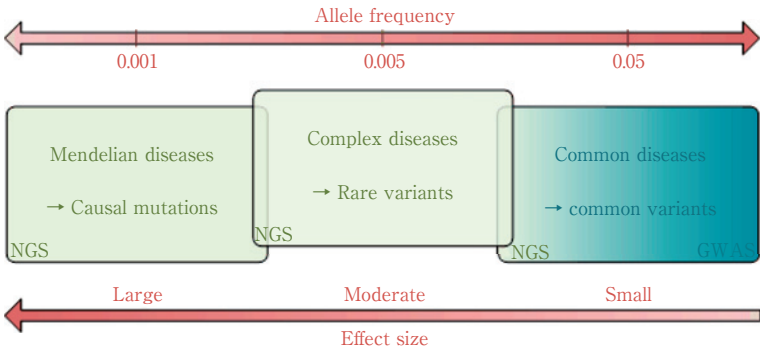

Fig. 3　Genetic approaches to identify variants of different effect sizes and frequencies.

more evidence supports the hypothesis of multiple rare variants for complex diseases such as Parkinson, Alzheimer disease or autism, so the ability to identify multiple genetic defects in a high throughput manner, as does NGS, will prove to be a valuable tool when studying complex diseases. The total of moderate effects from a series of variants with low frequency (usually below 2%), each contributing to an increased risk, may explain a large part of the inherited susceptibility for these complex diseases (Cohen, J.C. 2004; Mitsui, J. 2009; Fearnhead, N.S. 2004; Nejentsev, S. 2009; Ji, W. 2008). Because these rare variants have a low population frequency and probably bring forth only moderate effects they are not included in or their effect would not have been picked up in genome wide association studies. So the accurate detection of these rare variants may become pivotal in our understanding of complex disease development (Fig. 3).

Similarly, NGS provides us with the opportunity to search for additional genetic modifiers that may explain clinical heterogeneity (including differences in onset age or treatment response) within a family or within a group of unrelated carriers of the same mutation, on a genome wide or exome wide level.

NGS is also likely to replace mRNA arrays for gene expression. It can be used to sequence the mRNA directly, both sense and antisense, and again is non-biased in its approach. It will provide sequence for whatever is in the sample, not just what is on an expression chip. This sequence is called the transcriptome. Coupled with whole exome sequencing, it is potentially a powerful research tool.

Although we hear much discussion on whole genome sequencing, it is likely that whole exome sequencing is going to be the major tool for quite awhile. The reasons for this are practical; the exome represents only 1% of the total genome, thus bioinformatic and data management requirements are much smaller than for the whole genome. Further, in order to know what is abnormal, you need to understand normal. And most of our knowledge to date on normal sequence is centered on the exome, we are only just beginning to understand the rest of the genome, which is highly regulatory in its nature.

Of course, NGS can be used in many other applications and has indeed also gathered interest in different fields of the genetic research, such as mitochondrial studies, epigenetics or metagenomics; all of which might provide valuable information on disease development. For example, due to the small mitochondrial DNA size and thus the possible high level of coverage, NGS allows for the detection of heteroplasmy percentages when studying mitochondria, while the combination of NGS and epigenetics technology provides the opportunity to examine whole genome methylation. Also, sequencing the microbiome of certain human tissues can help us better understand the influence of the environment on ongoing biological processes in sickness and health.

In summary, whole exome sequencing and NGS is beginning to change medical practice. Not only does NGS enable screening of a possibly large, specific group of genes or region of interest in one experiment, it also allows unbiased detection on genome or exome level of novel causal variants and possible additional factors influencing onset or treatment outcome within one individual. This indicates that NGS is steering the medical field towards a genomic medicine or a more personalized medicine, in which NGS is the new major diagnostic tool, where treatment response can be predicted and risk algorithms for different diseases can be developed.

References

1) Moore GE. Cramming more components onto integrated circuits. Electronics 1965;38:114-117.
2) Cohen JC. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 2004;305:869-872.
3) Mitsui J. Mutations for Gaucher disease confer high susceptibility to Parkinson disease. Arch Neurol 2009;66:571-576.
4) Fearnhead NS. Multiple rare variants in different genes

account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci U S A 2004;101:15992-15997.

5) Nejentsev S. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009;324:387-389.

6) Ji W. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 2008;40:592-599.